MANAV RACHNA INTERNATIONAL UNIVERSITY

Short Synopsis

For

Ph. D. Programme 2013-14

Title: Design and Development of Protein Interactions Tool for Data Management in Systems Biology

MANAV RACHNA
vidyaparikshan

DEPARTMENT OF BIOTECHNOLOGY

FACULTY OF ENGINEERING & TECHNOLOGY

Submitted by:
Name: Rashmi Rameshwari
Signature: R. Rameshwari
Registration No.: 13/Ph.D/023
Email ID: rashmi.fet@mriu.edu.in
Contact No.: +91-9810039850

Supervisor:
Name: Dr. Shilpa S. Chapadgaonkar
Signature Shilpa
Designation: Assistant Professor
Email ID: shilpas.fet@mriu.edu.in
Contact No.: +91-9868052680

Joint Supervisor:
Name: Dr. T. V. Prasad
Signature:
Designation: Dean
Email ID: tvprasad2002@yahoo.com
Contact No.: +91-9493239815

ABSTRACT

Computer-aided models of biological networks are helpful in exploring different biological interactions. A variety of modeling tools have been developed till date to simulate biological interactions such as gene interactions, cellular physiology, and metabolic pathway. The metabolism of a cell consists of many interlinked reactions. Products of one reaction can catalyze the next. Thus, a cascade of networks is formed. Visualizing these networks and finding regulatory relationships between genes and their products is a challenging task. Complexity arises when large datasets is taken, as there is difficulty in finding individual interaction among entire proteome.

In this research problem, an attempt will be made to develop a tool, to simplify the analysis and interpretation of protein-protein interaction network by transforming the raw data into logically structured and visually tangible representations. This tool can build data-rich graphical representations that are color coded for Protein-protein interactions. It will be tested on disease dataset. This tool will be applicable in visualization of protein networks of many diseases. This tool will be equipped with many basic functionality to query the network; to visually integrate the network with expression profiles, phenotypes, and other molecular states and to link the network to databases of functional annotations.

Keywords: Biological Networks, Protein-protein Interaction, Gene interaction, Cellular physiology, Computational analyses.

Table of content

S. No.	Description	Page No.
1	Introduction	1
2	Literature Review	1
3	Description of Topic	5
4	Objectives of Research	6
5	Methodology to be adopted	6
6	Proposed/expected outcome of the research	8
7	References	9

1. Introduction

The basic building blocks of life, genes and their products (proteins), do not work independently but they interact with each other and form a very complex network. Gene networks represent the relationship between sets of genes that coordinate to achieve different tasks. Network models have facilitated a shift from the study of evolutionary conservation between individual gene and gene products towards the study of conservation at the level of pathways and complexes. With the advent of high-throughput microarray technologies, mRNA expression levels of tens of thousands of genes can now be measured simultaneously. This whole-genome-scale high throughput technology has revolutionized research in the life science arena since its invention (Lipshutz *et al.* 1995; Schena *et al.*1995). Large amount of microarray data has been generated since then and deposition of these data in public domains has made the study of gene network possible. Construction of gene networks from these experimental data will greatly facilitate cellular functional dissection at the molecular level. Gene networking study will advance the whole spectrum of biomedical research, from cancer research, toxicology, to drug discovery and disease prevention.

Many tools exist for visually exploring biological networks such as Patika (Demir *et al.* 2002), Cytoscape (Paul Shannon *et.al* 2003), Pathway Studio (Nikitin *et al.* 2003), VisANT 4.0 (Zhenjun Hu *et.al* 2013). These systems play a key role in the development of integrative biology, systems biology and integrative bioinformatics. The trend in the development of these tools is to go beyond 'static' representations of cellular state, towards a more dynamic model of cellular processes through the incorporation of gene expression data, sub cellular localization information and time-dependent behavior.

In the present work various tool as well as computational models have been considered.

2. Literature Survey

Protein interactions tools are based on various model and algorithm. Till date various models are available to design visualization tool. In this study few models have been considered that contain following features.

Model Survey:

There are many network visualization tools and a variety of computational models have been considered for reconstructing gene networks from gene expression data; such as, linear models (Van Someren *et al.* 2000; D'Haeseleer *et al.*, 2000; De Hoon, *et al.* 2003; Di Bernardo *et al.* 2004; De Jong *et al.* 2004; Deng *et al.* 2005; Chen *et al.* 2005;), Boolean network models (Huang, 1999; Kauffman, 1993; Shmulevich, *et al.*, 2002). Correlation network, (Steuer et al., 2002), Graphical Gaussian models (GGMs) (Opgen-Rhein and Strimmer, 2007). Among these network model, Bayesian network (BN) based approaches has been found to be better in showing relationship of gene interaction as well as protein interactions. The following reasons make it suitable:

(1) BNs can deal with the noises that are inherent in experimental measurements.

(2) BNs can handle missing data and permit the incomplete knowledge about the biological system.

(3) BNs are capable of integrating prior biological knowledge into the system.

Generally, a BN is a graphical representation of the dependence structure between multiple interacting quantities. This graphical representation is more commonly called a directed acyclic graph (DAG). The nodes or the vertices of the DAG represent the random variables in the network while the edges connecting the vertices represent the causal influence of one node on the other. Many structure learning methods have been proposed, which can be categorized as either conditional independence (CI) test-based methods or scoring-based methods. The CI-based methods analyze the dependence and independence relationships among variables and construct the networks that characterize these relationships (Geiger *et.al.* 1993; Meek 1995; de Campos and Huete 2000; Cheng *et al.* 2002).

Tool Survey

There are many genomic (gene-gene interaction) and proteomic (protein-protein) interactions network visualization tools and each meets some specific visualization method. In this study few tools have been considered that contains many of the important features

necessary for network visualization in a broad range of studies. The brief introduction of these tools is as follow:

The **PATIKA** tool (Demir *et al.* 2002) provides a world wide web based application, for accessing a central database containing pathway data from several sources. It provides high-quality visualizations using the Tom Sawyer Visualization software. It also supports SQL-like queries on node and edge properties. It is implemented using Java.

Pathway Studio (Nikitin *et al.* 2003) (formerly PathwayAssist) is a tool which uses Microsoft Windows application. It combines various features with better graphic user interface. It provides various display styles for visual attributes such as node color, size and shape, multi-user support, sub-cellular localization visualization and tight integration with several database including, KEGG (Kanehisa 2002), MINT (L Licata *et al.* 2012), GO (Ashburner *et al.* 2000), DIP (Xenarios *et al.,* 2000). The best part of this package is that it contains an SQL-like language that allows users to query the network. In comparison to other packages, this query language provides a much more flexible tool for quickly filtering nodes of interest from large networks.

Cytoscape (Shannon *et al.,* 2003) is another network visualization tool which supports many features such as standard and customizable network display styles, ability to import a large variety of interaction files, and zoomable network views. Cytoscape Release 2.3 supports networks with as many as 100K nodes and edges. One of the most positive aspects of Cytoscape is the large user and developer base. Cytoscape is a Java application whose source code is released under the Lesser General Public License (LGPL), At present more than 40 plugins are available for performing various task such as analyzing networks and generating networks from literature searches.

ProViz (Iragne *et al.,* 2005) leverages the power of the graph drawing package Tulip (David, 2001) for handling graphs containing millions of nodes and edges, while maintaining a guaranteed response time. Many plugins are currently available and shipped with the system including plugins for importing/exporting various network file formats, obtaining many different two and three-dimensional network layouts, computing various network metrics like connectivity and eccentricity and selecting sub graphs such as spanning sub trees and paths between nodes. Adds to these features the ability to import and export PSI-MI (Hermjakob *et*

al., 2001 and IntAct (Hermjakob *et al.*, 2004) data formats, and interfaces for exploring the GO (Ashburner *et al.*, 2000) and IntAct controlled vocabularies. There is functionality to define filters for large networks. Both ProViz and Tulip are implemented in C++ and can be installed in Windows, Linux or MacOSX.

BiologicalNetworks / PathSys (Baitaluk *et al*. 2006a) extends Cytoscape and includes functionality of Pathway Studio. BiologicalNetworks is a user-interface to PathSys a graph-based system for creating a combined database of biological pathways, gene regulatory networks and protein interaction maps. It integrates over 14 curated and publicly available data sources including BIND (Alfarano *et al*. 2005), GO (Ashburner *et al.*, 2000) and KEGG (Kanehisa, 2002) for eight representative organisms. BiologicalNetworks provides various features after importing expression data, users can apply sorting, normalization and clustering algorithms on the data and then create various tables, heat maps and network views of the data. It is implemented using Java Server Pages and is publicly available on the World Wide Web (WWW).

Graphle (Curtis Huttenhower *et al.2009)* is a visualization tool, used to search particular gene in existing biological interaction network. A gene information and graph is generated from the server. It is equipped with various filter to visualize biological network.

BIANA Interlog Prediction Server (Javier Garcia-Garcia et al. 2012) integrates several databases. Protein interaction is based on interlogs for a set of proteins provided by the user as input which can be entire proteome. This server predicts by pair to pair similarity to perform its inferences. It uses large datasets of experimentally identified PPIs. It also provides us information about clusters of orthologus genes.

VisANT 4.0 (Hu *et al.* 2013) not only provides network drawing but also provides convenient network analysis of diseases, therapies and drug target in network. VisANT 4.0 implements algorithms for analyzing node degrees, clusters, path lengths, network motifs and network randomizations. It facilitates meta-graph based solution with integrated disease and their therapies.

PINV: Protein Interaction Network Visualizer (Salazar et al. 2014) is an open source web application that facilitates that visualization of Protein interaction network through BioJS and uses D3 library to create graph layout. It provides three types of visualization of the data, such as network, circle and table.

IMP: Integrative Multispecies Prediction: (Aaron K Wong et al. 2015) is a web server which helps molecular biologist to interpret experimental results. Here role of candidate gene can be analyzed in large network. Human disease can be inferred directly in association with Online Mendelian Inheritance in Man.

DYNet (Ivan H.Goenawan et al. 2016), a cytoscape plugin developed to visualize multiple graph at a time. It provides editing of each graph and manages and synchronizes in real time via a central reference network. It also facilitates pairwise and multiple network comparision.

Research Gap

Literature survey of the recent work has revealed that :

- Most of the work focuses only on high-level networks low-level models and interactions are not utilized.
- Many tools uses only Microarray data to explore biological network. However a lot of experimental data has been explored using techniques other than microarray.

3. Description of Topic

3.1 Problem Identification

From literature it was found that due to complexity in proteomic data many problems arise in visualization. To remove these complexities an attempt will be made to develop a visualization tool which may illustrate the big picture represented by protein-protein interaction data, for qualitative assessment.

The problem undertaken for this research work is the development of an efficient visualization technique for protein-protein interactions. Till recently, the entire graph generated in protein interaction networks were in cyclic manner. But, in biological phenomenon cyclic as well as acyclic interactions occur. So in present work an attempt will be made to represent protein interactions in combination of cyclic as well as acyclic graph.

This tool would create a graphical representation of what is already "known" in literature and in molecular interaction repositories.

3.2 Analysis of the Problem

Presently entire protein interactome data cannot be visualized on a single platform. Huge amount of data is now being generated and the complexity of the interacting biological networks is being revealed. To explore the implications of biological network there is need to consider the interactome data sets. Protein data is not visible on one platform. There is need of efficient tool by which visualization of protein-protein interaction can be enhanced. By clicking one protein one can visualize entire interactions as well as its regulatory region.

3.3 Solution Strategy

For the above mentioned problem, the refinement in the existing model as well as development of improved algorithm for network graph generation is required. It is characterized by many parameters such as how proteins will interact, where they will interact, under what conditions they will interact, which is governed by many hypothesis and experiment.

4. Objectives of Research

1. To study network models and algorithms for integration of prior biological knowledge.
2. To design and develop protein interaction networks analysis tool for data integration.
3. To investigate network modularity, connectivity and proteins with particular roles in network structure using developed tool.

5. Methodology to be adopted

The methodology proposed to solve the above mentioned problem can be roughly divided into following steps:

For enhancement of existing work, an extensive survey on various bioinformatics tools related to biological path visualization and simulation will be carried out. Emphasis will be given on the enhancement of data integration and graph layout technique. Graphical browsers allow the user to examine all attributes on the currently selected nodes and edges of any given datasets.

Algorithm Designing

The present protein interaction network is visualized using cyclic graph. Therefore, to enhance the visualization, new algorithm development or refinement of the existing algorithms is proposed.

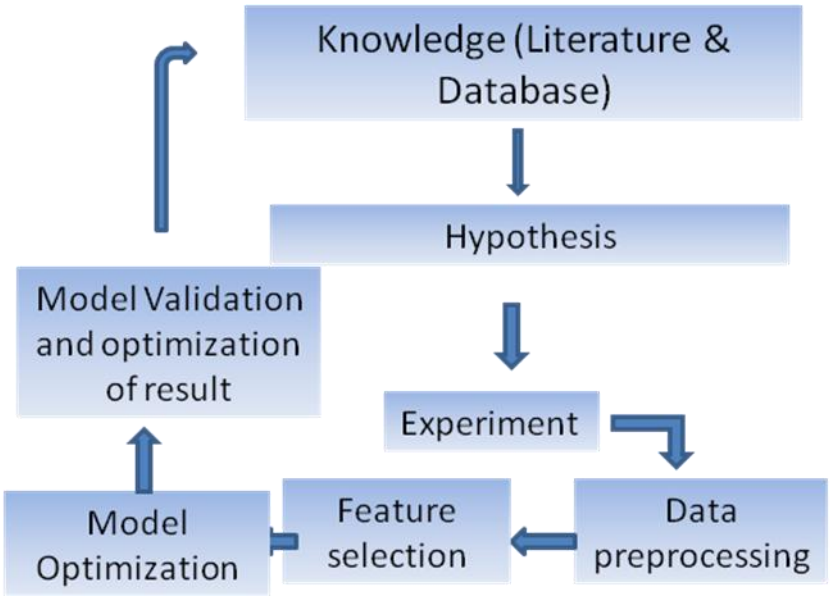

Figure: Systems Biology Cycle

Model Refinement

For better visualization, model refinement is required for existing work. Many network architectures such as correlation network, Bayesian networks, Graphical Gaussian models, Boolean networks, Module networks and many others are suggested by different researchers as per literature. Among these, selection of best model as per present work is required. Bayesian

Networks provide a very flexible framework for combining different types of data. It can also combine prior knowledge in the process of gene regulatory network inference to derive a suitable network structure. Beside this Bayesian networks has ability to avoid over fitting a model to training data and to handle incomplete noisy data as well as hidden variables.

Design of Tool for Biological network Visualization

For better network visualization an attempt will be made to design a tool based on the proposed algorithm and model.

The datasets of disease will be collected and tested on the developed tool. A comparative analysis will be carried out with existing tools to evaluate the performance of the newly developed tool.

6. Expected Outcome of the research

The main focus of this work is to make bioinformatics approach to visualize protein-protein interaction from given dataset. The proposed tool to be developed in this environment may attract large user communities, avoid duplication of effort and ultimately lead the way towards the holistic approach i.e. systems biology.

Significance:

With the help of this tool, a "Systems Biology" approach may be established by utilizing interdisciplinary resources. It has practical benefit in understanding of physiological and pathological processes of many diseases. The systems biology approach, with its combination of computational, experimental and observational enquiry, is highly relevant to drug discovery and the optimization of medical treatment regimes for individual patients. It will able to explore mechanism of many genetic diseases which sometimes occur due to irregularity in interaction mechanism for which drugs are yet to be designed. The span of any drug discovery directly affects its cost and time to the end user. If the span is reduced automatically, end users particularly the poor will be benefitted. With this gene profiling method a society free from disease can be assumed.

References

- Acid, S & De Campos, L, 2003, 'Searching for Bayesian network structures in the space of restricted acyclic partially directed graphs', *Journal Artificial. Intelligence*. Research, vol. 18, pp. 445–490.
- Baitaluk, M, Sedova, M, Ray, A., & Gupta, A, 2006, 'BiologicalNetworks: visualization and analysis tool for systems biology', *Nucleic Acids Research*, vol.34, pp. W466–471.
- Baitaluk, M, Qian, X, Godbole, S, Raval, A, Ray, A & Gupta, A, 2006, 'Pathsys: integrating molecular interaction graphs for systems biology', *BMC Bioinformatics*, vol.7, pp. 55-68.
- Demir, E, Babur, O, Dogrusoz, U, Gursoy, A, Nisanci, G & Cetin-Atalay, R, 2002, 'PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways', *Bioinformatics*, vol.18, pp. 996–1003.
- Friedman, N, Geiger D & Goldszmidt M, 1997, 'Bayesian network classifiers', *Machine Learning*, vol. 29, pp. 131–163.
- Goenawan IH, Bryan K, Lynn DJ, 2016, 'DyNet: Visualization and analysis of dynamic molecular interaction networks', Bioinformatics, btw187.
- Garcia-Garcia J, Schleker S, Seetharaman JK , Oliva B, 2012,'BIPS: BIANA Interolog Prediction Server. A tool for protein–protein interaction inference', Nucleic Acids Research, Vol. 40, pp. W147–W151.
- Goutsias, J & Lee, NH, 2007, 'Computational and experimental approaches for modeling gene regulatory networks', *Current Pharmaceutical Design*, vol.13, pp. 1415–1436.
- Hu, Z, Chang, Y, Wang, Y, Huang, C, Liu,Y, Tian, F, Granger, B & DeLisi, C, 2013, 'VisANT: Integrative network platform to connect genes, drugs, diseases and therapies', *Nucleic Acids Research*, vol. 33, pp.W225–W321.

- Huttenhower C, Mehmood OS, Troyanskaya OG, 2009, 'Graphle: Interactive exploration of large, dense graphs',BMC Bioinformatics, 10:417.
- Hartemink AJ, Gifford DK, Jaakkola TS & Young RA, 2001, 'Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks, *Pacific Symposium on Biocomputing*, pp. 422–433.
- Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D & Apweiler R, 2004, 'IntAct: an open source molecular interaction database', *Nucleic Acids Research*, vol. 32, pp. D452–D455.
- Iragne F, Nikolski M, Mathieu B, Auber D, & Sherman D, 2005, 'ProViz: protein interaction visualization and exploration', *Bioinformatics*, vol. 21, 272–274.
- Ikin A, Riveros C, Moscato P & Mendes A, 2010, 'The Gene Interaction Miner: a new tool for data mining contextual information for protein-protein interaction analysis', *Bioinformatics*, vol.26(2), pp. 283-287.
- Martin S, Zhang, Z, Martino, A & Faulon, JL, 2007. 'Boolean dynamics of genetic regulatory networks inferred from microarray time series data', *Bioinformatics*, vol.23 (7), pp. 866–874.
- Nikitin, A, Egorov, S, Daraselia, N, Mazo, I, 2003, 'Pathway studio – the analysis and navigation of molecular networks', *Bioinformatics*, vol. 19(16), pp. 2155–2157.
- Rameshwari R & Prasad TV, 2010, 'A Survey of Various Protein-Protein Interaction Tools', *Proceedings of the National Conference on Advances in knowledge*', pp. 229-238.
- Salazar GA, Meintjes A, Mazandu GK, Rapanoel HA, Akinola RA,Mulder NJ, 2014 'A web-based protein interaction network visualizer', BMC Bioinformatics, **15**:129.
- Shannon P, Markiel, A, Ozier, O, Baliga, NS, Wang, JT, Ramage, D, Amin, N, Schwikowski, B, & Ideker, T, 2003, 'Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*earch, vol. 13, pp. 2498–2504.
- Schwikowski B, Uetz P, Fields S, 2000, 'A network of protein-protein interactions in yeast', *Nature. Biotechnology*, vol. 18, pp. 1257–1261.

- Van Someren, EP, Wessels, LFA and Reinders, MJT, 2000, 'Linear Modelling of genetic networks from experimental data in intelligent systems for Molecular Biology', *ISMB 2000*, San Diego, CA, pp. 355–366.
- Van Berlo, RJP, Van Someren, EP & Reinders, MJT, 2003, 'Studying the conditions for learning dynamic Bayesian networks to discover genetic regulatory networks', *Simullation: Transactions of Society for Modeling and Simulation International*, vol. 79 (12), pp. 689–702.
- Wong AK, Krishnan A, Yao V, Tadych A,Troyanskaya OG, 2015, 'IMP 2.0: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks', Nucleic Acids Research, Vol. 43, pp. W128–W133.
- Natal AW & Van Riel, 2006, 'Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments', *Briefings in Bioinformatics*, vol. 7(4), pp.364–374.
- Xenarios, I, Rice, DW, Salwinski, L, Baron, K.B, Edward, M.M & Eisenberg, D, 2000, DIP: the database of interacting proteins. Nucleic Acids Res., vol. 28(1), pp. 289–291.
- Zanzoni, A, Palazzi, LM, Quondam, M, Ausiello, G, Citterich, HM, Cesareni, G, (2002) MINT: a molecular interaction database. *FEBS Lett*., vol. 513, pp. 135–140.